WORKSHOP 1

ANA CRISTINA QUINTERO CARPINTERO

2226286

Javier Alejandro Vergara Zorrilla

ETL
UNIVERSIDAD AUTÓNOMA DE OCCIDENTE

AGOSTO - 28 – 2024

WHAT IS EXPECTED?

I expect you to get the CSV file and create an application to migrate the data to a relational database. Also, you will display that data from the database in graphical visualizations - remember, the data should be stored in a database and your reports should come from the database, not the CSV file.

The visualizations I expect are:
Hires by technology (pie chart)
Hires by year (horizontal bar chart)
Hires by seniority (bar chart)
Hires by country over years (US, Brazil, Colombia, and Ecuador only) (multi-line chart)

Technologies
We expect you to use in this challenge:
Python
Jupiter Notebook
Database (Postgres)
Diagram

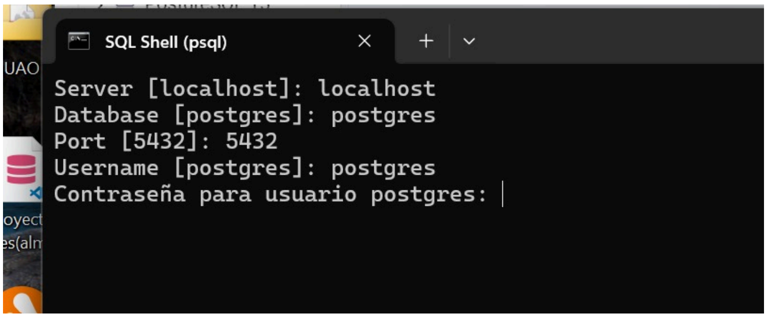
Workshop -001: Data Engineer 2
Data
I have 50,000 rows of data about candidates. The fields we'll be using are:
First Name
Last Name
Email
Country
Application Date
Year of Experience
Seniority
Technology
Code Challenge Score
Technical Interview

Remember that I consider a candidate to be HIRED when they have both scores greater than or equal to 7 - you need to apply this logic to get the right information. How you'll handle this data is up to you.

And remember that all the data here is totally random - we used a public library to generate random information.

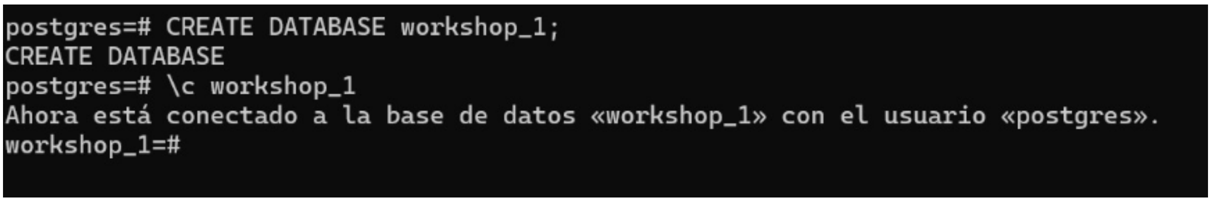
In this document, the entire development of workshop 1 will be presented.

It is important to mention that the Postgres database was chosen, therefore, the connection to Postgres is created.



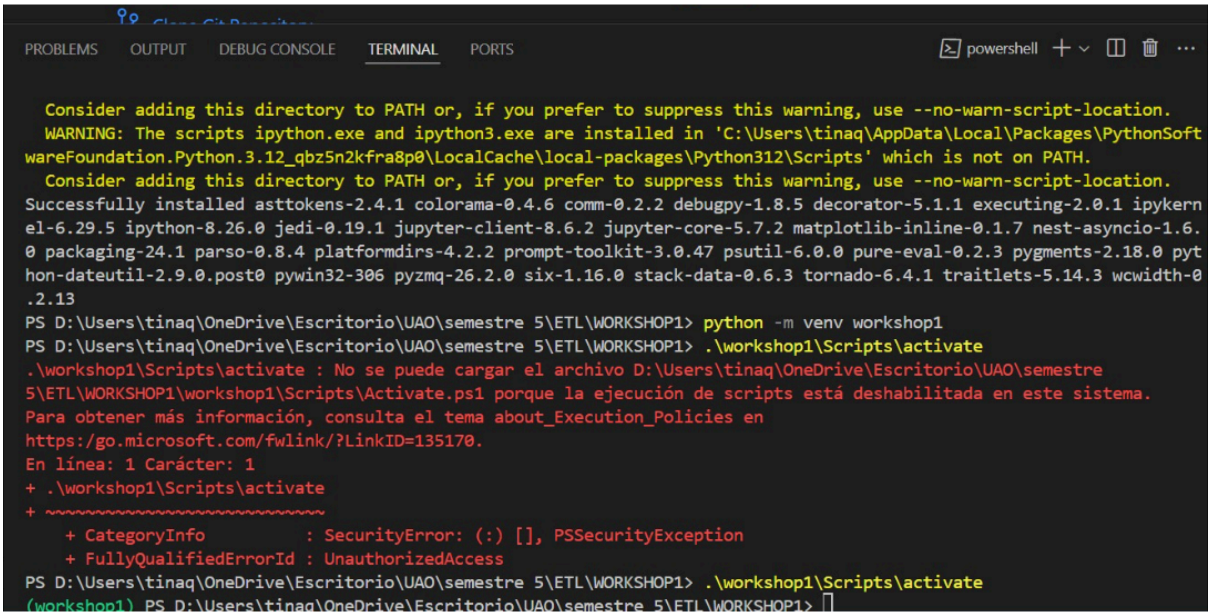
```
Server [localhost]: localhost
Database [postgres]: postgres
Port [5432]: 5432
Username [postgres]: postgres
Contraseña para usuario postgres:
```

We proceed to create the database.



```
postgres=# CREATE DATABASE workshop_1;
CREATE DATABASE
postgres=# \c workshop_1
Ahora está conectado a la base de datos «workshop_1» con el usuario «postgres».
workshop_1=#
```

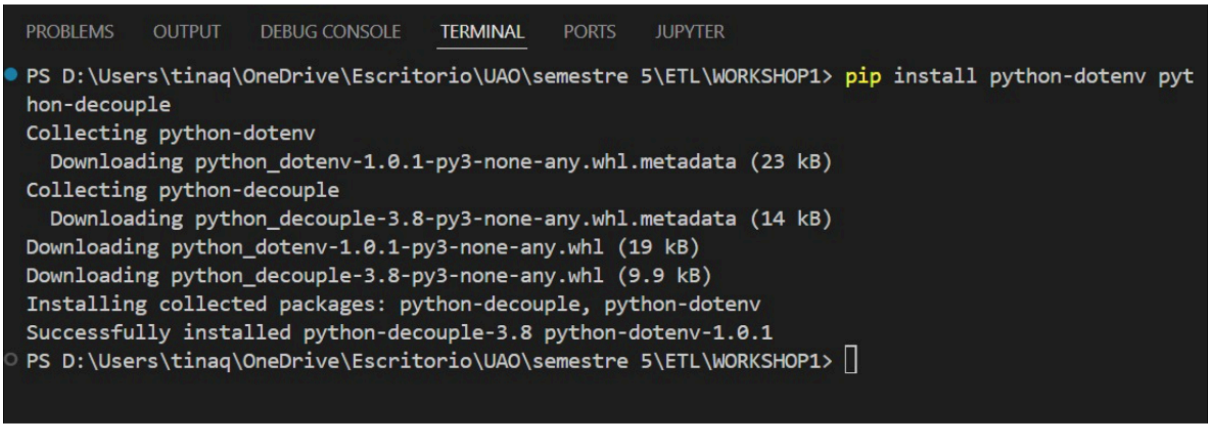
We proceed to create the kernel in the virtual environment



```
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
  WARNING: The scripts ipython.exe and ipython3.exe are installed in 'C:\Users\tinaq\AppData\Local\Packages\PythonSoft
wareFoundation.Python.3.12_qbz5n2kfra8p0\LocalCache\local-packages\Python312\Scripts' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed asttokens-2.4.1 colorama-0.4.6 comm-0.2.2 debugpy-1.8.5 decorator-5.1.1 executing-2.0.1 ipykern
el-6.29.5 ipython-8.26.0 jedi-0.19.1 jupyter-client-8.6.2 jupyter-core-5.7.2 matplotlib-inline-0.1.7 nest-asyncio-1.6.
0 packaging-24.1 parso-0.8.4 platformdirs-4.2.2 prompt-toolkit-3.0.47 psutil-6.0.0 pure-eval-0.2.3 pygments-2.18.0 pyt
hon-dateutil-2.9.0.post0 pywin32-306 pyzmq-26.2.0 six-1.16.0 stack-data-0.6.3 tornado-6.4.1 traitlets-5.14.3 wcwidth-0
.2.13
PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1> python -m venv workshop1
PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1> .\workshop1\Scripts\activate
.\workshop1\Scripts\activate : No se puede cargar el archivo D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre
5\ETL\WORKSHOP1\workshop1\Scripts\Activate.ps1 porque la ejecución de scripts está deshabilitada en este sistema.
Para obtener más información, consulta el tema about_Execution_Policies en
https:/go.microsoft.com/fwlink/?LinkID=135170.
En línea: 1 Carácter: 1
+ .\workshop1\Scripts\activate
+ ~~~~~~~~~~~~~~~~~~~~~~~~~~~~
    + CategoryInfo          : SecurityError: (:) [], PSSecurityException
    + FullyQualifiedErrorId : UnauthorizedAccess
PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1> .\workshop1\Scripts\activate
(workshop1) PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1>
```

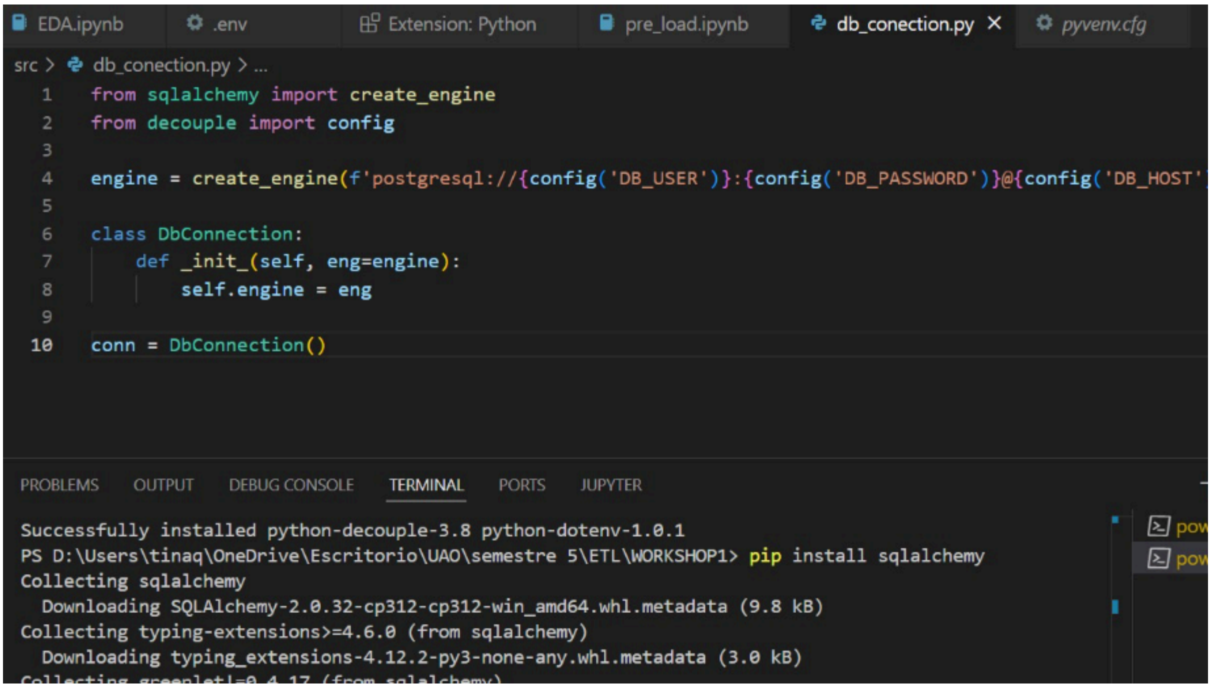
(workshop1) PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1> pip install ipykernel

We proceed to install the libraries in the environment variable



```
PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1> pip install python-dotenv pyt
hon-decouple
Collecting python-dotenv
  Downloading python_dotenv-1.0.1-py3-none-any.whl.metadata (23 kB)
Collecting python-decouple
  Downloading python_decouple-3.8-py3-none-any.whl.metadata (14 kB)
Downloading python_dotenv-1.0.1-py3-none-any.whl (19 kB)
Downloading python_decouple-3.8-py3-none-any.whl (9.9 kB)
Installing collected packages: python-decouple, python-dotenv
Successfully installed python-decouple-3.8 python-dotenv-1.0.1
PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1>
```

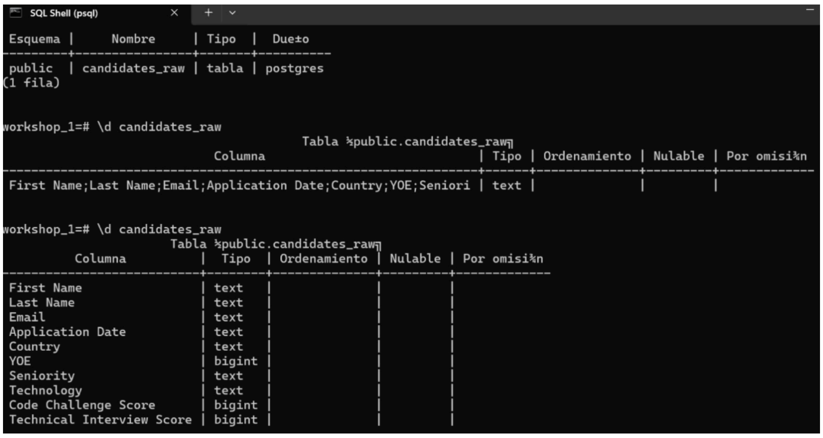
We proceed to create the connection to the database



```python
from sqlalchemy import create_engine
from decouple import config

engine = create_engine(f'postgresql://{config('DB_USER')}:{config('DB_PASSWORD')}@{config('DB_HOST')

class DbConnection:
    def _init_(self, eng=engine):
        self.engine = eng

conn = DbConnection()
```

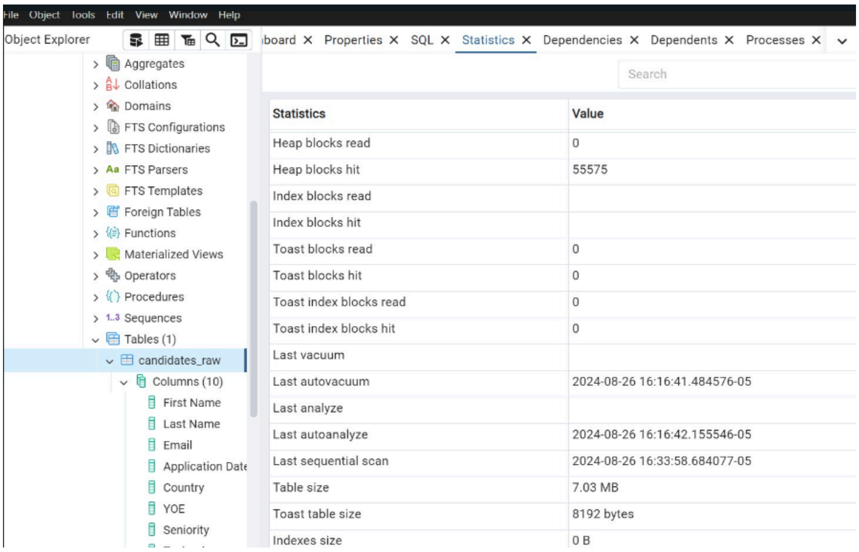
```
Successfully installed python-decouple-3.8 python-dotenv-1.0.1
PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1> pip install sqlalchemy
Collecting sqlalchemy
  Downloading SQLAlchemy-2.0.32-cp312-cp312-win_amd64.whl.metadata (9.8 kB)
Collecting typing-extensions>=4.6.0 (from sqlalchemy)
  Downloading typing_extensions-4.12.2-py3-none-any.whl.metadata (3.0 kB)
Collecting greenlet!=0.4.17 (from sqlalchemy)
```

We proceed to verify that the database is previously loaded



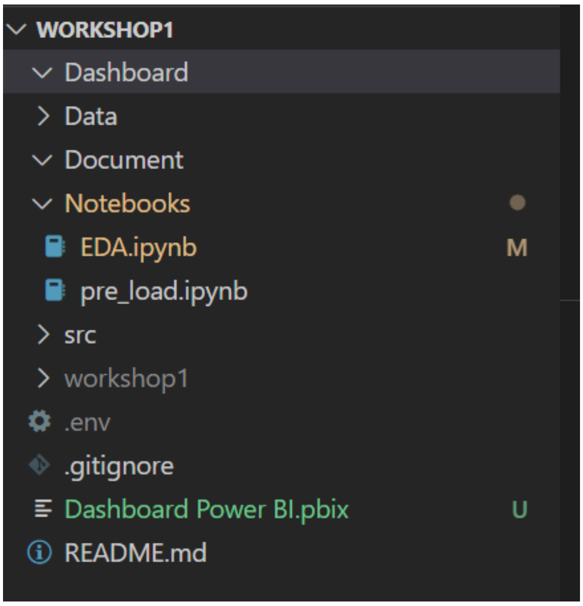
```
 Esquema |     Nombre     | Tipo  |  Due±o
---------+----------------+-------+----------
 public  | candidates_raw | tabla | postgres
(1 fila)


workshop_1=# \d candidates_raw
                                   Tabla ½public.candidates_raw╗
                             Columna                              | Tipo | Ordenamiento | Nulable | Por omisi¾n
------------------------------------------------------------------+------+--------------+---------+-------------
 First Name;Last Name;Email;Application Date;Country;YOE;Seniori | text |              |         |


workshop_1=# \d candidates_raw
                    Tabla ½public.candidates_raw╗
         Columna           |  Tipo  | Ordenamiento | Nulable | Por omisi¾n
---------------------------+--------+--------------+---------+-------------
 First Name                | text   |              |         |
 Last Name                 | text   |              |         |
 Email                     | text   |              |         |
 Application Date          | text   |              |         |
 Country                   | text   |              |         |
 YOE                       | bigint |              |         |
 Seniority                 | text   |              |         |
 Technology                | text   |              |         |
 Code Challenge Score      | bigint |              |         |
 Technical Interview Score | bigint |              |         |
```

We validate that the dirty database is previously loaded.

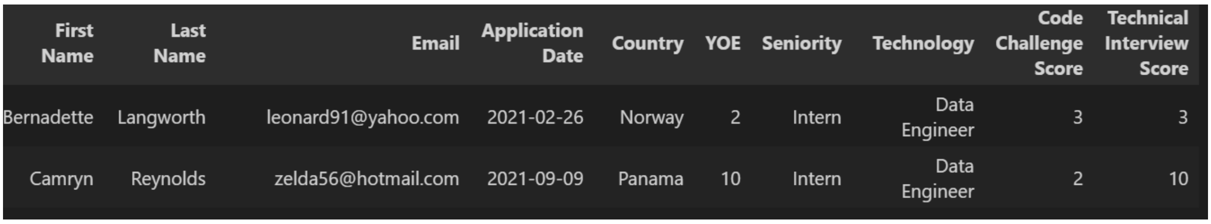


Statistics	Value
Heap blocks read	0
Heap blocks hit	55575
Index blocks read	
Index blocks hit	
Toast blocks read	0
Toast blocks hit	0
Toast index blocks read	0
Toast index blocks hit	0
Last vacuum	
Last autovacuum	2024-08-26 16:16:41.484576-05
Last analyze	
Last autoanalyze	2024-08-26 16:16:42.155546-05
Last sequential scan	2024-08-26 16:33:58.684077-05
Table size	7.03 MB
Toast table size	8192 bytes
Indexes size	0 B

We visualize the folders in the project:

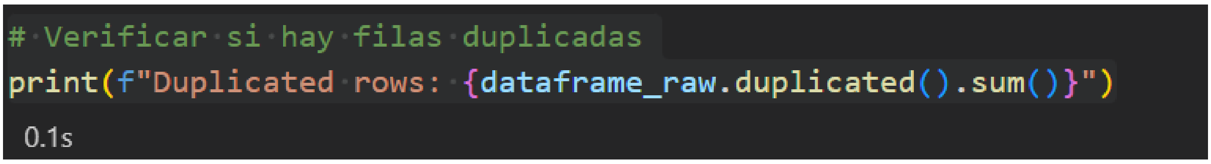


Then we start the analysis process:
It is important to install the libraries and read the database



First Name	Last Name	Email	Application Date	Country	YOE	Seniority	Technology	Code Challenge Score	Technical Interview Score
Bernadette	Langworth	leonard91@yahoo.com	2021-02-26	Norway	2	Intern	Data Engineer	3	3
Camryn	Reynolds	zelda56@hotmail.com	2021-09-09	Panama	10	Intern	Data Engineer	2	10

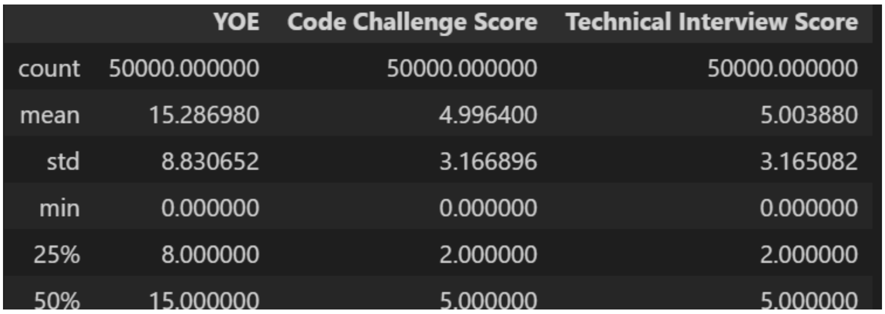
A review is made of how the data is found:



```
# Verificar si hay filas duplicadas
print(f"Duplicated rows: {dataframe_raw.duplicated().sum()}")
```

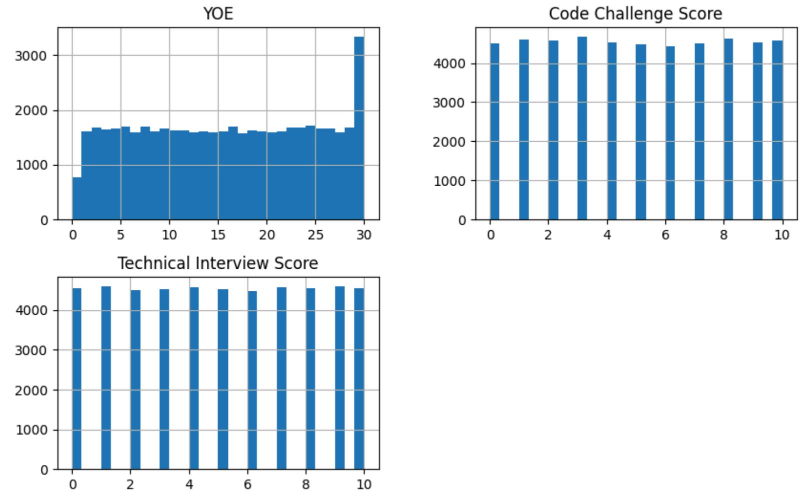
```
0.1s
```

To do this, we review the distribution of the data and obtain the following results:



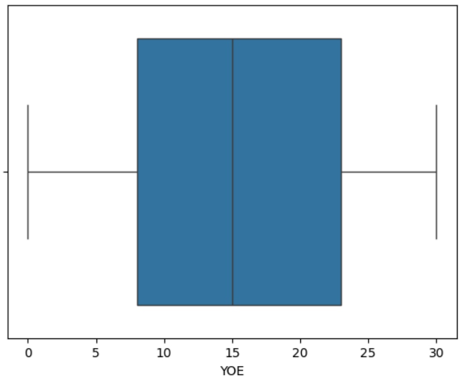
	YOE	Code Challenge Score	Technical Interview Score
count	50000.000000	50000.000000	50000.000000
mean	15.286980	4.996400	5.003880
std	8.830652	3.166896	3.165082
min	0.000000	0.000000	0.000000
25%	8.000000	2.000000	2.000000
50%	15.000000	5.000000	5.000000

Subsequently, the exploratory data analysis process is carried out and therefore we obtain the following conclusions:

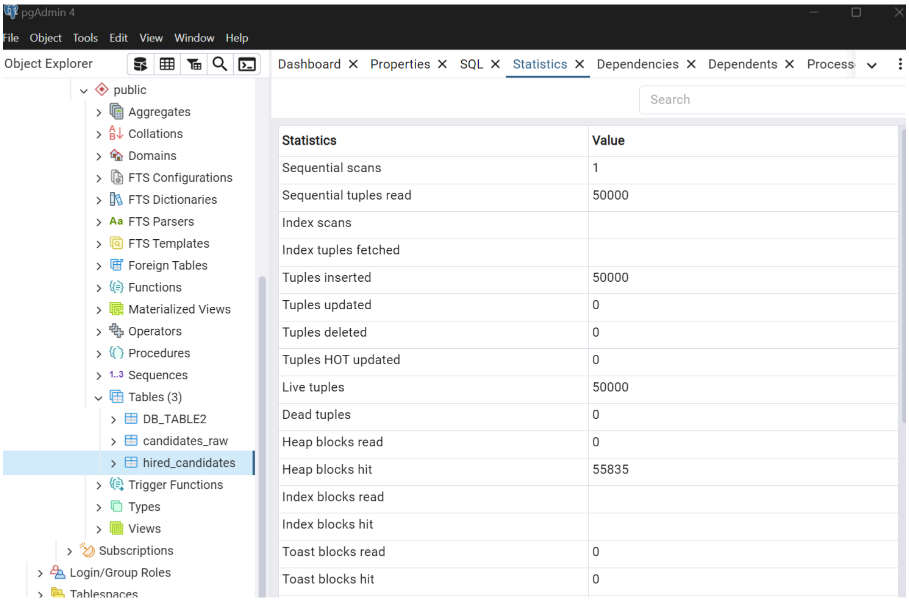


It is observed that there is a balanced distribution in the scores of the technical evaluations, with many candidates located in the middle range.

According to the graph, we can see that the years of experience variable shows a significant diversity in terms of the experience levels of the candidates.
It is detected that there is a large concentration of candidates with 30 years of experience and this indicates that there may be a concentration among the data, which leads us to look for outliers.



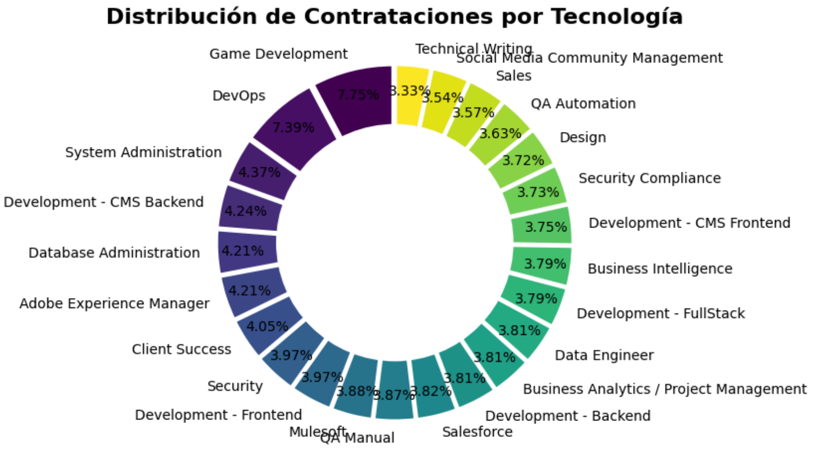
After cleaning and EDA is saved in the database (hired_candidates)



Statistics	Value
Sequential scans	1
Sequential tuples read	50000
Index scans	
Index tuples fetched	
Tuples inserted	50000
Tuples updated	0
Tuples deleted	0
Tuples HOT updated	0
Live tuples	50000
Dead tuples	0
Heap blocks read	0
Heap blocks hit	55835
Index blocks read	
Index blocks hit	
Toast blocks read	0
Toast blocks hit	0

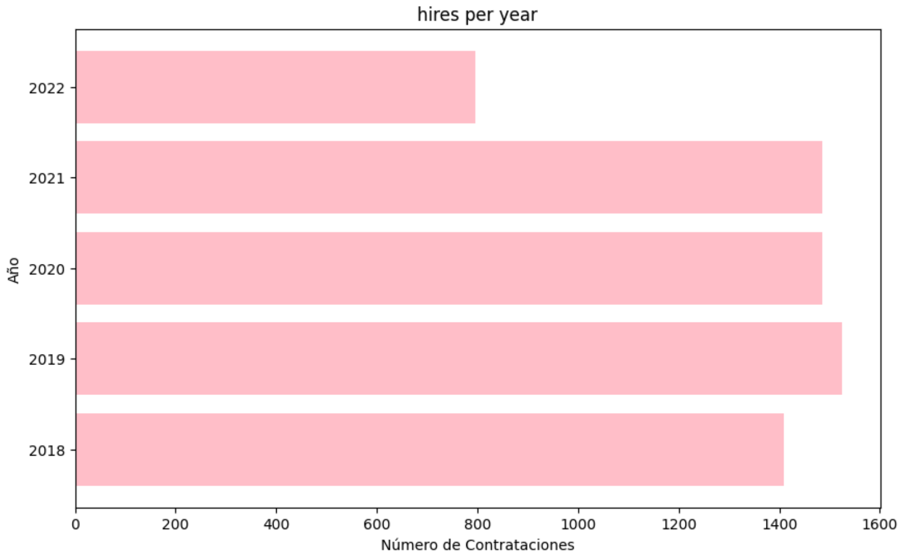
Finally, since we have the clean table stored in our database, we begin to create the requested graphics.

- Hiring by technology (pie chart)



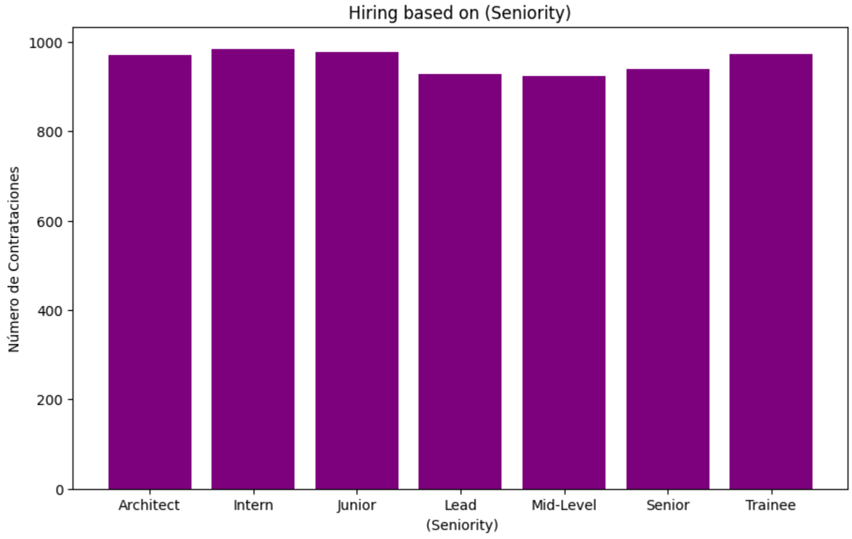
In the graph above we can see how the hiring of prospective employees is distributed across various fields in the technology area.

- Hires per year (horizontal bar chart)



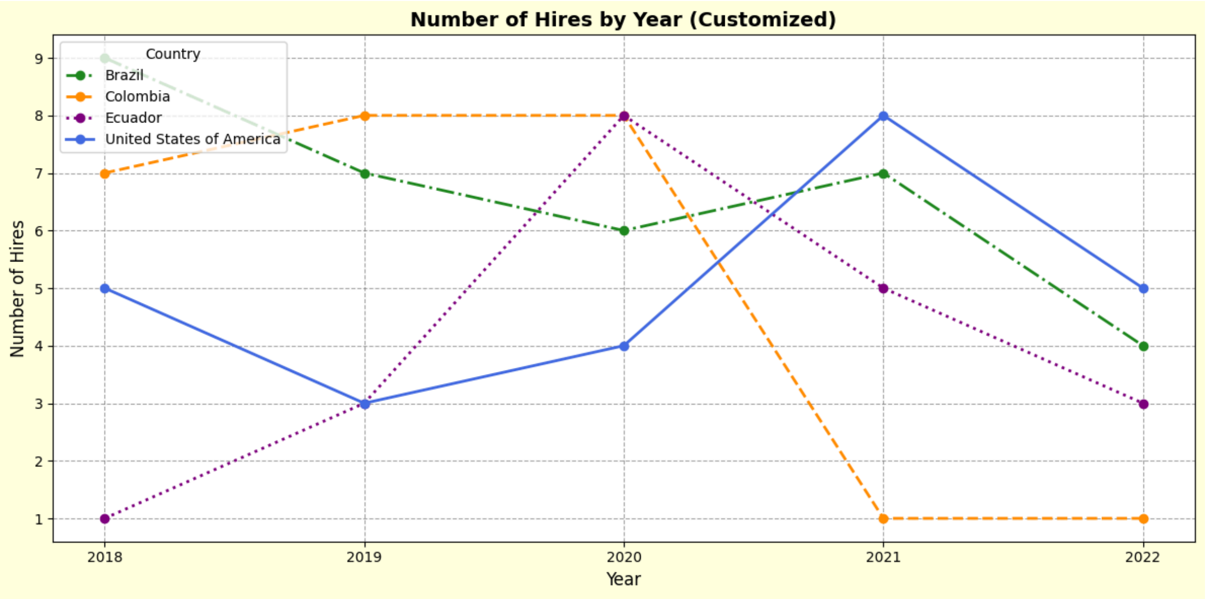
In the graph above we can see how the hiring of prospects is distributed over the various years.

- Hires by seniority (bar chart)



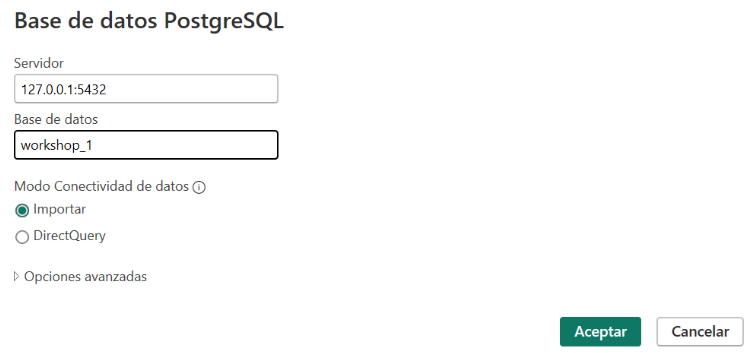
In the graph above we can see how the hiring of prospective employees is distributed by seniority.

- Hires by country over the years (US, Brazil, Colombia and Ecuador only) (multi-line chart)



In the graph above we can see how the hiring of prospects is distributed by selected countries over the years.

Finally, the database connection to the Power BI tool is made.



Base de datos PostgreSQL

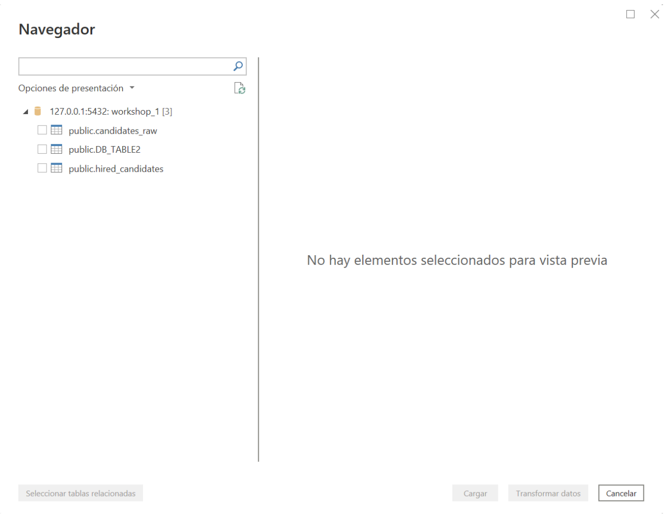
Servidor: 127.0.0.1:5432

Base de datos: workshop_1

Modo Conectividad de datos

● Importar

○ DirectQuery



Navegador

127.0.0.1:5432: workshop_1 [3]

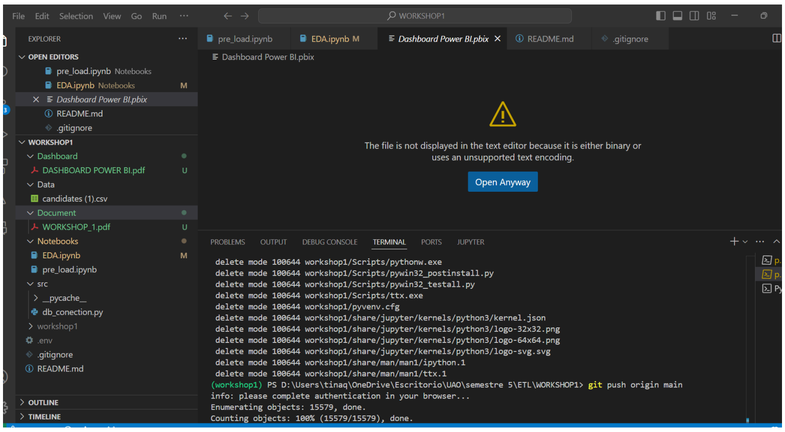
- □ public.candidates_raw
- □ public.DB_TABLE2
- □ public.hired_candidates

No hay elementos seleccionados para vista previa

The following business questions were asked to implement in Power BI.

Which countries have contributed the highest number of candidates?
How does the years of experience (YOE) vary based on the seniority level of the candidates?
Is there any correlation between the country of origin of the candidates and the technologies they are specialized in?
What is the average duration between the application date and the hiring date?
Relationship between the year of application with the year of the technical interview
Relationship between Code Challenge Score and Technical Interview

Se sube desde el Visual Studio, todo el trabajo a GitHub



The file is not displayed in the text editor because it is either binary or uses an unsupported text encoding.

```
 delete mode 100644 workshop1/Scripts/pythonw.exe
 delete mode 100644 workshop1/Scripts/pywin32_postinstall.py
 delete mode 100644 workshop1/Scripts/pywin32_testall.py
 delete mode 100644 workshop1/Scripts/ttx.exe
 delete mode 100644 workshop1/pyvenv.cfg
 delete mode 100644 workshop1/share/jupyter/kernels/python3/kernel.json
 delete mode 100644 workshop1/share/jupyter/kernels/python3/logo-32x32.png
 delete mode 100644 workshop1/share/jupyter/kernels/python3/logo-64x64.png
 delete mode 100644 workshop1/share/jupyter/kernels/python3/logo-svg.svg
 delete mode 100644 workshop1/share/man/man1/ipython.1
 delete mode 100644 workshop1/share/man/man1/ttx.1
(workshop1) PS D:\Users\tinaq\OneDrive\Escritorio\UAO\semestre 5\ETL\WORKSHOP1> git push origin main
info: please complete authentication in your browser...
Enumerating objects: 15579, done.
Counting objects: 100% (15579/15579), done.
```

Los comandos usados fueron:

git init

git rm --cached -r workshop1

git add .

git commit -m

git push origin main